Weather and Climate Extremes 32 (2021) 100318

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: http://www.elsevier.com/locate/wace

Clustering bivariate dependencies of compound precipitation and wind extremes over Great Britain and Ireland

Edoardo Vignotto [a,*], Sebastian Engelke [a], Jakob Zscheischler [b,c,d]

[a] Research Center for Statistics, University of Geneva, Geneva, Switzerland
[b] Climate and Environmental Physics, University of Bern, Sidlerstrasse 5, 3012, Bern, Switzerland
[c] Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[d] Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

ARTICLE INFO

Keywords:
Extremes
Spatial clustering
Compound events

ABSTRACT

Identifying hidden spatial patterns that define sub-regions characterized by a similar behaviour is a central topic in statistical climatology. This task, often called regionalization, is helpful for recognizing areas in which the variables under consideration have a similar stochastic distribution and thus, potentially, for reducing the dimensionality of the data. Many examples for regionalization are available, spanning from hydrology to weather and climate science. However, the majority of regionalization techniques focuses on the spatial clustering of a single variable of interest and is often not tailored to extremes. Extreme events often have severe impacts, which can be amplified when co-occurring with extremes in other variables. Given the importance of characterizing compound extreme events at the regional scale, here we develop an algorithm that identifies homogeneous spatial sub-regions that are characterized by a common bivariate dependence structure in the tails of a bivariate distribution. In particular, we use a novel non-parametric divergence able to capture the similarities and differences in the tail behaviour of bivariate distributions as the core of our clustering procedure. We apply the approach to identify homogeneous regions that exhibit similar likelihood of compound precipitation and wind extremes in Great Britain and Ireland.

1. Introduction

Statistical modelling of climate extremes such as heatwaves, droughts, heavy precipitation, storms, and floods is of major societal interest (Field et al., 2012). Extreme events may occur in many different situations, often with dramatic consequences. A better understanding of the underlying processes of extreme events can help to mitigate potentially severe impacts on society.

Many large impacts are caused by compound events, for instance the concurrent occurrence of multiple and possibly interdependent hazards (Leonard et al., 2014; Zscheischler et al., 2018, 2020). Such co-occurring extreme events can lead to larger impacts compared to univariate extremes (Zscheischler et al., 2014; Ribeiro et al., 2020). In these cases, univariate techniques that analyze the effect of each hazard separately may be misleading and underestimate the true overall risk (Zscheischler and Seneviratne, 2017). Nonetheless, so far much of the analysis of extreme events has focused on individual extremes of a single variable (Field et al., 2012).

One main line of research in statistical climatology is to identify sub-regions characterized by a similar behaviour of the variables of interest. This is of fundamental importance for two main reasons (Grimaldi et al., 2016). Firstly, this allows climate scientists to gain a deep knowledge of the phenomenon under consideration and its different spatial effects, eventually facilitating the implementation of spatially specialized measures that minimize potential adverse socio-economic impacts. Secondly, determining regions that can be roughly considered statistically similar to a given location of interest with respect to their distributional behaviour can improve the estimation of specific quantities at this location when records are scarce (Pappadà et al., 2018). This regionalization operation, as it is commonly called in hydrology and other disciplines with heterogeneous datasets, allows to retrieve precise estimates of high quantiles of a given variable even when only few data are at disposal at the considered location, pooling together records from other sites recognized as roughly equivalent (Asadi et al., 2018). Regionalization also helps to pool highly non-stationary and heterogeneous data along locations with a similar tail behaviour, for instance to

* Corresponding author.
E-mail address: edoardo.vignotto@unige.ch (E. Vignotto).

https://doi.org/10.1016/j.wace.2021.100318
Received 31 July 2020; Received in revised form 14 November 2020; Accepted 4 March 2021
Available online 15 March 2021

model the spatial structure of rainfall extremes (Saunders et al., 2020).

Many clustering algorithms have been proposed to summarize spatial complex climatological datasets and to capture some features of their underlying hidden patterns. However, most of these approaches focus on a single variable and compare univariate distributions at different locations. Though some clustering approaches have been extended to include multiple variables (Zscheischler et al., 2012), ultimately they are still based on metrics between univariate distributions. Within the clustering approaches based on comparison between univariate distributions, over the recent years, a number of approaches have been suggested focusing on extremes. For example, weekly precipitation maxima in France have been analyzed by Bernard et al. (2013), who developed a clustering algorithm based on a proper distance justified by extreme value theory. The same approach is followed in Bador et al. (2015) to asses the bias of climate model simulations of temperature maxima over Europe. Durante et al. (2015) used a similar approach based on tail dependence to cluster time series data. Pappadà et al. (2018) proposed a distance metric based on copulas to cluster similarities between flood peaks. Regarding the use of clustering techniques to reduce the statistical uncertainty associated with high return level estimates at a site with scanty data, Asadi et al. (2018) developed an optimal procedure to identify similar locations that can be used to enhance the precision of those estimates for river discharge measurements.

The multivariate case, in which two or more different variables are considered at the same time, is more challenging. With respect to compound extremes, many methods based on extreme value theory have been proposed in recent years to study the multivariate distribution of two or more variables of interest given that one, or both, are extreme. A widely used approach is to disentangle the modelling of the marginal distribution and the residual dependence structure, following a copula approach (Kolev et al., 2006). Another popular approach is the conditional one, that studies the behaviour of a group of variable given a specific variable being large (Jonathan et al., 2014; Gouldby et al., 2017). Note that in this latter case, not all the considered variables need to lay in the tail region since the only one explicitly modeled as extreme is the conditioning variable (Heffernan and Tawn, 2004). Nonetheless, applications in climate science have been limited so far. Most of the above-mentioned extreme value models are based on parametric assumptions that presume a specific analytical form of the dependence structure and thus may produce a strong bias in the estimates if this assumption is wrong. Moreover, they usually do not allow a comparison of the bivariate dependence structures of the considered hazards at multiple locations, and therefore they do not permit to evaluate the similarity between two given sites with respect to the co-occurrence of extremes.

Building on this prior work, here we rely on the theoretical foundations of extreme value theory and use a non-parametric divergence measure for the tail behaviour of multivariate distributions to quantify the similarity in the structure of bivariate tail dependence. We demonstrate the applicability of this divergence measure using a widely used clustering approach, the *K*-medoids algorithm, to cluster regions in Ireland and Great Britain according to the dependence structure in the extremes of precipitation and wind speed. Compound precipitation and wind extremes can have dramatic impacts, such as human fatalities, impaired critical infrastructure and economic losses (Liberato, 2014; Raveh-Rubin and Wernli, 2015; Martius et al., 2016). For this reason, identifying sub-regions that are characterized by a similar extremal dependence structure may help design differential measures to mitigate impacts. Our primary goal is to offer with this application a first proof of concept of the merits and limitations of our methodology, with the aim of introducing it and facilitating its use.

The paper is organized as follows. Section 2 describes the data and methods used in our application and introduces the non-parametric divergence at the core of our clustering procedures. Section 3 illustrates the application of the approach to real world observations and discusses the obtained results before we present the main conclusions in Section 4.

2. Data and methods

2.1. Data

We downloaded daily precipitation sums and wind speed maxima over the Great Britain and Ireland extracted from the ERA5 dataset (Copernicus Climate Change Service (C3S), 2017) on a spatial resolution of 0.25° on a regular grid. This state-of-the-art reanalysis product is build with an updated physical weather model and data assimilation process compared to the previous ERA Interim (Dee et al., 2011).

We consider weekly sums of daily precipitation and weekly averages of daily wind speed maxima in winter (from November to March) at $N = 677$ locations in Great Britain and Ireland from January 1st, 1979 to December 31st, 2018. We choose a weekly temporal scale because precipitation and wind extremes can be linked through storms with a lag of several days due to persistent weather patterns (Bengtsson et al., 2009). With this event definition we focus on the risk posed by high wind speeds and abundant rainfalls occurring in a short amount of time even if they may not be caused from the same extratropical cyclone. For simplicity and following Bernard et al. (2013), we do not consider weeks with no precipitation.

Extremes in precipitation and wind speed often co-occur and we are interested in understanding the spatial variability of this relationship, which is of interest when studying weather systems associated with extreme precipitation and extreme winds. The goal is to identify regions for which the extremal dependence structure between these two variables shows a similar behaviour.

2.2. Characterizing dependence in the extremes

We briefly introduce here two fundamental concepts related to the characterization of the dependence structure of extreme compound events, namely the concept of asymptotic dependence and asymptotic independence (cf., Ledford and Tawn, 1997; Poon et al., 2004). Two variables X_1 and X_2 with cumulative distribution functions F_1 and F_2 respectively are said to be asymptotically dependent if

$$X = \lim_{q \to 1} \mathbb{P}(F_1(X_1) > q | F_2(X_2) > q) \tag{1}$$

$$= \lim_{q \to 1} \frac{\mathbb{P}(F_1(X_1) > q, F_2(X_2) > q)}{1 - q} \in (0, 1] \tag{2}$$

and asymptotic independent otherwise (i.e., if $\chi = 0$). The coefficient χ is called extremal correlation and represents, loosely speaking, the probability of one variable being 4 extreme given that the other one is extreme. Note that two variables can be dependent but asymptotically independent. For instance, in the case of a bivariate Gaussian distribution with correlation $\rho \in [-1, 1)$ we always have $\chi = 0$ (Sibuya, 1960). In this case the coefficient χ is clearly not informative since it is not able to distinguish the different strengths of associations between X_1 and X_2 for different correlation coefficients ρ. Under asymptotic independence, another coefficient, the residual tail dependence coefficient $\overline{\chi}$, describes the strength of the tail relationship (Ledford and Tawn, 1996). Indeed,

$$\overline{\chi} = \lim_{q \to 1} \frac{2 \log \mathbb{P}(F_1(X_1) > q)}{\log \mathbb{P}(F_1(X_1) > q, F_2(X_2) > q)} - 1 \tag{3}$$

$$= \lim_{q \to 1} \frac{2 \log(1 - q)}{\log \mathbb{P}(F_1(X_1) > q, F_2(X_2) > q)} - 1 \in [-1, 1] \tag{4}$$

is equal to 1 for asymptotically dependent variables, while for asymptotically independent variables its value indicates if X_1 and X_2 are positively ($\overline{\chi} > 0$) or negatively ($\overline{\chi} < 0$) associated in their extremes. In

the example of the bivariate Gaussian distribution, we have that $\overline{\chi} = \rho$. Thus, the pair of coefficients $(\chi, \overline{\chi})$ summarizes the tail dependence structure of X_1 and X_2. Finally, note that both χ and $\overline{\chi}$ are symmetric with respect to X_1 and X_2.

Since both the χ and $\overline{\chi}$ coefficients are defined as a limit value, a usual way to analyze the behaviour of a bivariate tail dependence structure between two variables is to compute their empirical estimates for different threshold levels q and then visually inspect their behaviour as $q \to 1$. Confidence intervals for a fixed q can be easily computed by bootstrapping. We will follow this approach in Section 3.

2.3. A divergence between bivariate extremal dependence structures

While χ and $\overline{\chi}$ are useful summary statistics on the extremal dependence structure between two variable X_1 and X_2, they do not permit to assess directly the level of similarity between the extremal behaviour of two bivariate random variables, say $\boldsymbol{X}^{(1)} = (X_1^{(1)}, X_2^{(1)})$ and $\boldsymbol{X}^{(2)} = (X_1^{(2)}, X_2^{(2)})$. For example, a $\chi^{(1)}$ can be computed between extreme precipitation and strong winds from one dataset, e.g., based on observations, and compared to a $\chi^{(2)}$ for a second dataset, e.g., a climate model simulation. But it would also be very convenient to have a single number to tell us if the extremal dependence between these two bivariate random vectors are different, and if so, by how much. Engelke et al. (2019) proposed to use a non-parametric dissimilarity measure for this purpose. They use the so-called Kullback–Leibler divergence, which is very popular in signal processing and other fields. This approach generalizes the concept of the coefficients χ and $\overline{\chi}$ and allows to compute the dissimilarity between the tail dependence structures between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ in a more complete manner. The divergence has been used by Zscheischler et al. (2021) to assess whether dependence structures between precipitation and wind extremes differ across different datasets. Here we apply the approach to identify coherent sub-regions characterized by similar extremal dependence structures.

We will quickly introduce the divergence between extremal dependence structures, similar to (Zscheischler et al., 2021). For two bivariate distributions $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, the divergence is only defined in the tail of the distributions after normalizing the marginal distributions to standard Pareto distributions. For a univariate random variable X with distribution function F, this normalization is done empirically by the operation $1/\{1 - \widehat{F}(X)\}$. Fig. 1 shows weekly sums of winter precipitation and daily wind speed maxima for two selected locations both on the original scale and with margins normalized to standard Pareto distributions (Fig. 1a–b and c-d respectively). This is a common transformation in extreme value statistics that highlights the tail region and it is required to define the divergence as explained in the following. A risk function computed on the Pareto scale $r : \mathbb{R}^2 \to \mathbb{R}$ is used to define which areas in the bivariate distributions are considered extreme. There are different choices for the risk function. Taking the sum function $r(\boldsymbol{x}) = x_1 + x_2$, $\boldsymbol{x} = (x_1, x_2)$ or the maximum function $r(\boldsymbol{x}) = \max(x_1, x_2)$, $\boldsymbol{x} = (x_1, x_2)$ give similar result. In this way, two new univariate variables $R^{(1)} = r(\boldsymbol{X}^{(1)})$ and $R^{(2)} = r(\boldsymbol{X}^{(2)})$ are defined. We consider those points as extremes for which the variable $R^{(j)}$ exceeds a given high quantile $q_u^{(j)}$ corresponding to an high exceedance probability $u \in (0, 1)$, $j = 1, 2$. Varying the threshold $q_u^{(j)}$ alters the extremal region of interest. Using the sum risk function, for each of the two bivariate distributions, the set

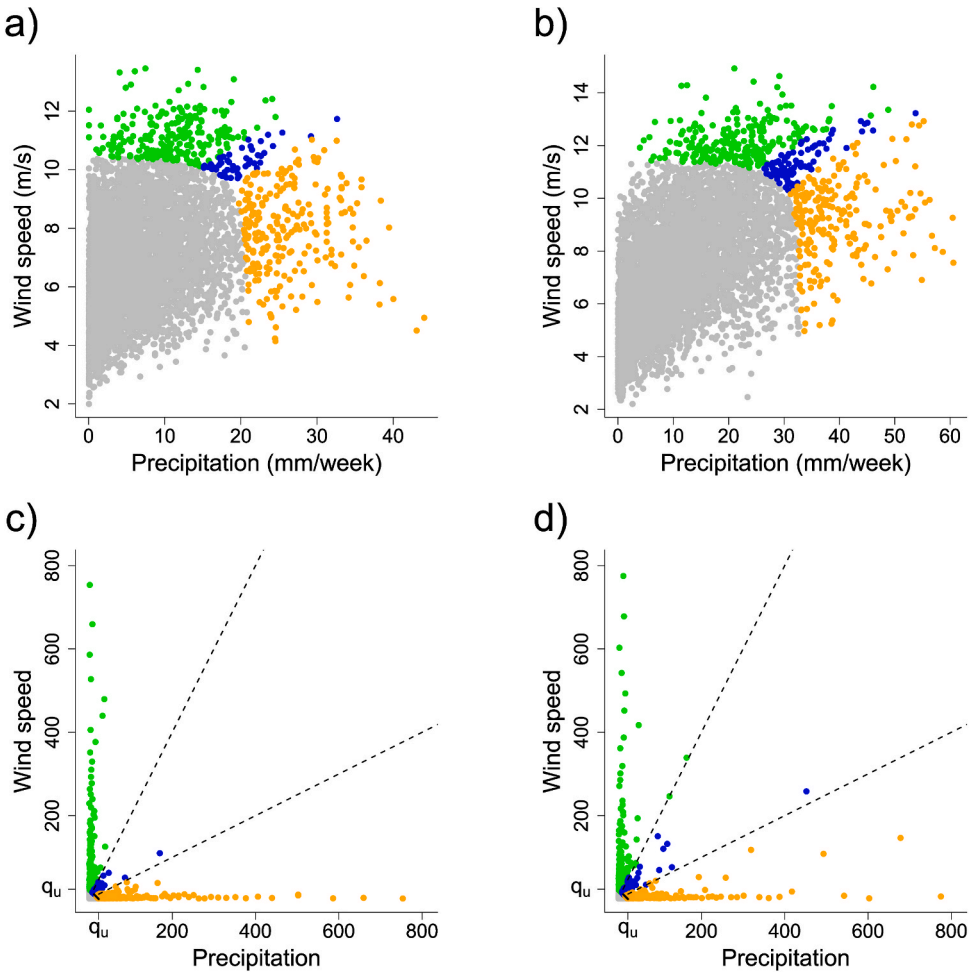

Fig. 1. An illustration of the partitioning used in the divergence measure in (6). Weekly total precipitation and daily wind speed maxima in winter for two selected locations (a–b), and after their marginal distributions were normalized to standard Pareto (c–d). To compute the divergence, the tails of the bivariate distributions of interest need to be separated into W disjoint subsets, here represented with different colors and separated by dashed lines in the bottom plots (c–d) using $W = 3$ and $u = 0.9$ for the sum risk function. The ratios between the probabilities associated with these subsets are then used as a measure of dissimilarity between the tail dependence structures of the two bivariate distributions, see main text for a more detailed explanation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

$\{R^{(j)} > q_u^{(j)}\}$, $j = 1, 2$, contains the extreme points. This set is now partitioned into a fixed number W of disjoint sets $A_1^{(j)}, \ldots, A_W^{(j)}$ as shown in Fig. 1 (for $W = 3$ and $u = 0.9$). For the maximum risk function the data are divided in $W = 3$ different sets, one containing the co-occurring extremes and the other two containing data for which only one of the variables is extreme, see the illustration in Fig. S1.

Suppose now that we have two random samples $\boldsymbol{X}_1^{(1)}, \ldots, \boldsymbol{X}_n^{(1)}$, and $\boldsymbol{X}_1^{(2)}, \ldots, \boldsymbol{X}_n^{(2)}$, from the distributions $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. The empirical proportions of data points belonging to each of the above sets $A_w^{(j)}$ is computed as

$$\widehat{p}_w^{(j)} = \frac{\#\{i : \boldsymbol{X}_i^{(j)} \in A_w^{(j)}\}}{\#\{i : r(\boldsymbol{X}_i^{(j)}) > q_u^{(j)}\}}, \quad w = 1, \ldots, W; j = 1, 2; i = 1, \ldots, n. \tag{5}$$

The difference between the extremal behaviours of the two distributions can then be measured as the Kullback–Leibler divergence between the two multinomial distributions defined through these proportions, that is,

$$d_{12} = D(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}) = \frac{1}{2} \sum_{w=1}^{W} \left(\left(\widehat{p}_w^{(1)} - \widehat{p}_w^{(2)} \right) \log \left(\widehat{p}_w^{(1)} \Big/ \widehat{p}_w^{(2)} \right) \right). \tag{6}$$

The divergence in (6) is a natural way to look at different extremal dependence structures for both asymptotically dependent and independent data. This divergence, in addition, is symmetric and since it is a non-parametric statistic it does not require additional model assumptions. The definition of the divergence also underlines how it generalizes the χ and $\overline{\chi}$ coefficients. The latter only computes the number of points in sets where both margins exceed the same threshold, that is, $\{\boldsymbol{x} \in \mathbb{R}^2 : x_1 > q, x_2 > q\}$; see (1) and (3). The statistic in (6) takes into account the number of points in all the sets $A_w^{(j)}$ and is therefore able to capture potentially very complex dependence structure much better.

Regarding the free parameter W for the sum risk function, we note that, while many sets would permit to capture more subtle differences between the compared extremal dependence structures, a too high number W could results in many sets containing few data points, resulting in an undefined divergence d_{12}. Thus, the choice of this free parameter presents a classical bias-variance trade-off that may require a sensitivity analysis of the results.

2.4. The K-medoids algorithm

Provided a pairwise divergence, many algorithms are available to cluster objects into homogeneous groups. Here we use the bivariate divergence d_{ij} from (6) characterizing differences in the tail dependence between location i and location j as defined in Section 2.3. We consider as clustering method the K-medoids algorithm, that we outline below. The goal of the procedure, given N objects or data points $x_1, \ldots, x_N$, is to group them into K clusters in order to minimize a loss function, usually some kind of homogeneity measure. This function, together with the rules that define to which cluster each object belongs completely defines a clustering algorithm. The input of the algorithm is thus the matrix $\{d_{ij}\}_{i,j=1,\ldots,N}$ consisting of the divergences between any two data points and the number K of partitions in which to divide the N objects of interest. The output of the algorithm is a list that describes to which cluster each data point belongs.

The K-medoids algorithm was proposed in Kaufman and Rousseeuw (2009) as a more robust alternative to the K-means algorithm. The algorithm can be described with the following steps:

1 Fix a number K of clusters, and for each $k = 1, \ldots, K$, randomly choose a point $x_{i(k)}$ as initial center, called medoid.
2 Form K clusters by assigning every point $x_1, \ldots, x_n$ to its closest medoid (measured with the divergences d_{ij}).
3 For each cluster $C_k = 1, \ldots, K$, find the new medoid $x_{i(k)} \in C_k$ for which the total intracluster divergence $x_{i(k)} = \underset{x_i \in C_k}{\operatorname{argmin}} \sum_{x_j \in C_k} d_{ij}$ is minimized.
4 If at least one medoid has changed, then go back to point 2, otherwise end the algorithm.

In other words, the K-medoids algorithm tries to find a configuration of medoids that minimizes the divergences between each data point and its cluster center. Note that with the K-medoids algorithm each cluster center is one of the N objects given as the input, in our case one of the locations of interest. To minimize the influence of the random initial point in step 1 a possibility is to reinitialize the algorithm many times and retain the partitioning that minimize some measure of goodness of fit such as the (average) silhouette coefficient introduced below.

The only hyper-parameter of the K-medoids algorithm is the number of clusters K to consider. A useful tool in this regard is the so-called silhouette coefficient (Rousseeuw, 1987). This measure compares the divergences between objects belonging to the same cluster and objects belonging to different clusters. Intuitively, a good partitioning should put near data point into one cluster and distant data points into different clusters. For each observation i, the silhouette coefficient $s(i)$ is defined as following. Denote with $a(i)$ the average divergence between i and the other observations in its cluster, with $b(i, k)$ the average divergence between i and the observations in cluster k and with $c(i)$ the minimum of $b(i, k)$ with respect to k. Then, the silhouette coefficient is defined as

$$s(i) = \frac{c(i) - a(i)}{\max(a(i), c(i))}.$$

Data points with $s(i)$ near to 1 are well clustered, while data points with $s(i)$ near to 0 are badly clustered.

3. Results and discussion

We focus on the tail dependence between compound precipitation and wind extremes. More specifically, at each location $j = 1, \ldots, N$, we examine the bivariate vector $\boldsymbol{X}^{(j)} = (X_1^{(j)}, X_2^{(j)})$, where $X_1^{(j)}$ and $X_2^{(j)}$ are the distributions of weekly total precipitations and weekly averages of daily wind speed maxima in winter, respectively. Note that we do not model explicitly the spatial dependence structure of the univariate variables X_1 and X_2. This analysis can be done modelling the spatial dependence structure of X_1 or X_2 as a function of the distance between locations (Wadsworth and Tawn, 2012) or as a function of one or more covariates (Winter et al., 2016). Otherwise, univariate clustering techniques for extremes are also available (Bador et al., 2015; Rohrbeck and Tawn, 2020). For example, Rohrbeck and Tawn (2020) propose a Bayesian spatial clustering technique of the extremal behaviour of a single variable of interest and apply it to daily precipitations in Norway and river flows in UK. We make this decision to focus on the novelty of our approach, i.e., the ability to cluster together areas characterized by a similar bivariate tail behaviour of two different variables without using prior knowledge on their spatial location.

We apply the clustering procedure to these bivariate distributions with divergence matrix $\{d_{ij}\}_{i,j=1,\ldots,N}$, where d_{ij} is the divergence measure for extremal dependence defined in (6). If not stated otherwise, we use $W = 3$ sets following Zscheischler et al. (2021) and an exceedance probability of $u = 0.9$ as parameters in the definition of the divergence (see Section 2.3). In this section, we present and discuss the results for the sum risk function. The results are not sensitive to the choice of the risk function (sum or maximum) and different choices of the number of partitioning sets W (see Supplementary Material).

We first perform an exploratory analysis of the extremal dependence structure of the considered variables. The estimated χ coefficient (with $q = 0.9$) between both variables suggest that the western region of Great Britain and Ireland is characterized by a stronger extremal dependence between weekly total precipitation and weekly averages of wind speed maxima (Fig. 2a). This spatial dependence structure is coherent with previous studies on precipitation and wind speed correlation over the

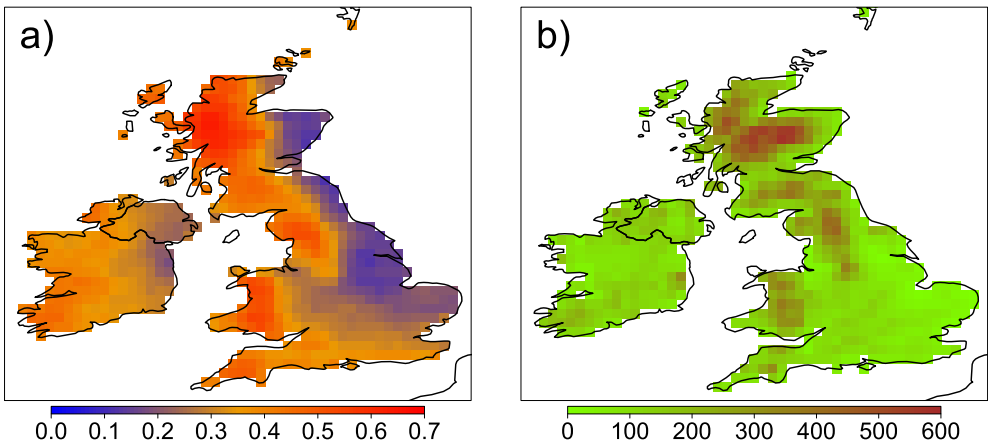

Fig. 2. (a) Grid-based tail dependence coefficient χ between weekly total precipitations and averages of daily wind speed maxima in winter ($q = 0.9$). (b) Altitude [m] map of Great Britain and Ireland.

considered area and is somewhat related to orography (Fig. 2b). Heavy precipitation in the western part of Great Britain and Ireland are observed in periods associated with a strong North Atlantic jet stream causing strong westerly wind anomalies (Baker et al., 2018) bringing moist air to the western flanks of the mountains as well as extratropical cyclones (Mailier et al., 2006; Hendry et al., 2019). The North Atlantic jet stream has less effect on the eastern zone due to the so-called rain shadow effect (Weston and Roy, 1994; Fowler et al., 2005; Svensson et al., 2015) for which areas located on the eastern side of mountain regions are effected by relatively lower precipitations due to westerly wind. In the east, high precipitation anomalies are usually associated with a low pressure field centered over the UK resembling the Eastern Atlantic pattern and characterized by a less correlated easterly wind speed anomalies (Baker et al., 2018). Overall, the results are also consistent with Martius et al. (2016), who found lower co-occurrence rates of compound precipitation and wind extremes in the lee of mountains.

After this exploratory analysis, we apply K-medoids clustering algorithm described in Section 2.4 with the divergence metric d_{ij} in (6). The silhouette coefficients obtained with different values of K suggest that the best partitioning is found for $K = 3$, with similar performance also for $K = 2$ and $K = 4$ (Fig. 3). Thus, we report the partitions obtained for $K = 2, 3, 4$ in Fig. 4. In all these cases, the obtained partitions reflect the exploratory results based on the χ coefficient (Fig. 2a) in combination with the elevation above sea level (Fig. 2b). Our algorithm is thus capable to identify a predefined number of sub-regions characterized by different extremal dependence strength and high spatial coherency without using any spatial constrains. The identification of regions with similar dependence structure in the extremes may help, for example, to build better regional forecasting models by defining areas with a similar stochastic behaviour where the same model can be applied (Baker et al., 2018).

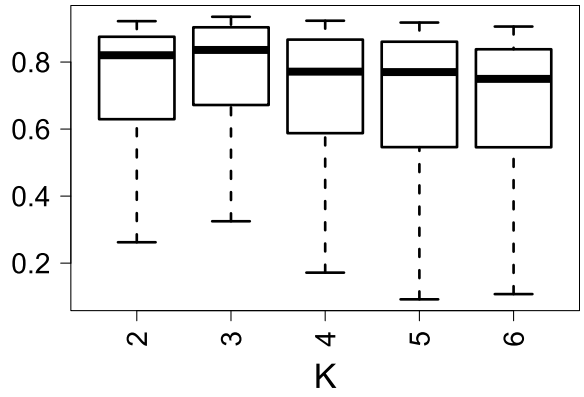

Fig. 3. Boxplots of the silhouette coefficients for different values of K. Thick lines indicate the median, boxes the interquartile range and whiskers the full range of the distribution.

In Fig. 5 we illustrate the bivariate distributions of the medoids, i.e., the cluster centers, identified through the clustering procedure with $K = 3$. The three medoids show different dependence structures, from a mild dependence for the first (a) to a stronger dependence for the third one (c). We further explore these differences with a particular focus on the tail region in Fig. 6. The estimated χ and $\overline{\chi}$ coefficients for these distributions are plotted as a function of the threshold level from $q = 0.6$ to $q = 0.99$, which is a typical exploratory plot in multivariate extreme value statistics. The colors correspond to colors in the bivariate distributions of the medoids shown in Fig. 5. Note that the coefficient χ is the limit as $q \to 1$, but in practice one has to estimate a version with $q < 1$. The interest is indeed not only in the limit, but also in the speed with which it converges to zero in the case of asymptotic independence. That is why one usually plots estimates for different values of q as exploratory data analysis. The same kind of reasoning holds also for $\overline{\chi}$.

The curves for the χ coefficient for all the three medoids tend to 0 as $q \to 1$ but with different velocities (Fig. 6a), indicating asymptotic independence. Overall this behaviour indicates that the lower the threshold for the extremes are, the more dependent are the extremes between both variables, as it is often the case. $\overline{\chi}$ provides further information on the dependence in the extremes, showing rather stable and distinct non-zero values for all three medoids (Fig. 6b). Specifically, the medoid depicted in gold is characterized by weak dependence in the extremes, especially for quantile levels q close to 1. The medoids for the other two clusters shown in red and purple of this figure have significantly stronger dependence in the extremes. This can already be seen in the data shown in the scatter plots, which are less spread out for extreme values of precipitation and wind speed (Fig. 5). Finally, looking at the confidence bands, the three medoids are associated with curves that are significantly different from one another for almost every quantile level, thus dividing the dependence space into distinct groups even when considering uncertainties. Plotting confidence bands for $\widehat{\chi}$ and $\widehat{\overline{\chi}}$ could thus give an indication of the magnitude of the differences in the tail dependence structure of the medoids.

Thus, the medoids offer a useful summary representation of the stochastic behaviour of the different clusters and this type of *a posteriori* analysis can serve as an effective way to gain deeper insights into the difference in extremal dependence between the clusters, which can then help construct regional models (Saunders et al., 2020).

Overall the results confirm widely known patterns of co-occurring wind and precipitation extremes in the Great Britain and Ireland (e.g.

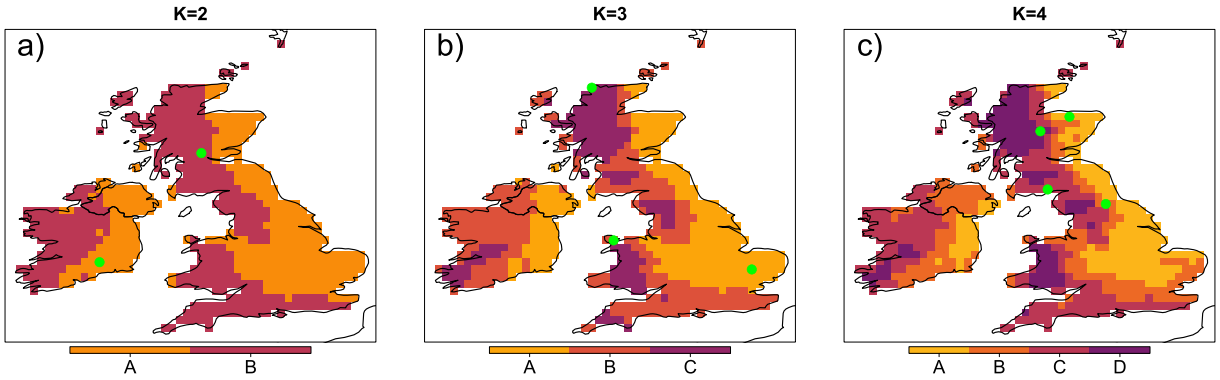

Fig. 4. Partitions obtained with different values of K ($K = 2$, K= 3 and $K = 4$ for a), b) and c), respectively). Locations that are medoids for the respective cluster are highlighted with green dots. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

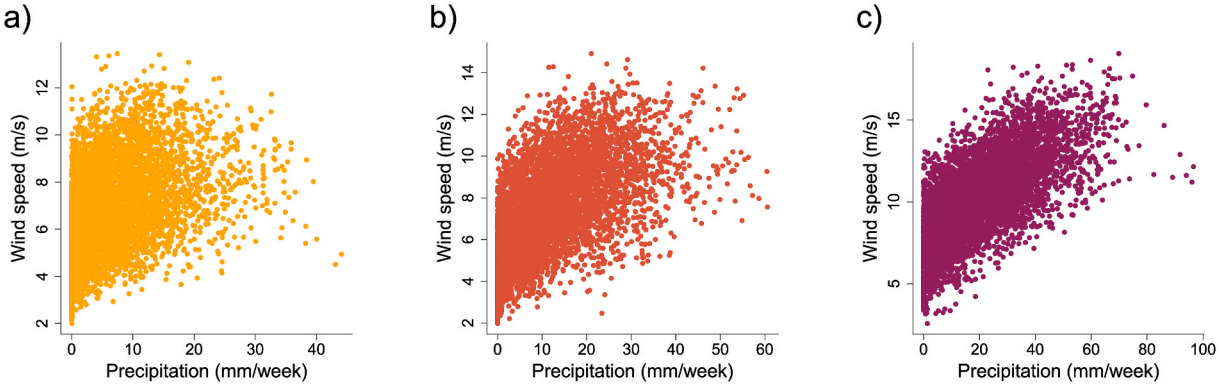

Fig. 5. Bivariate distributions of the medoids identified by the clustering algorithm with $K = 3$ classes, highlighted by green dots in Fig. 4b. The dependence increases from a) to c). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

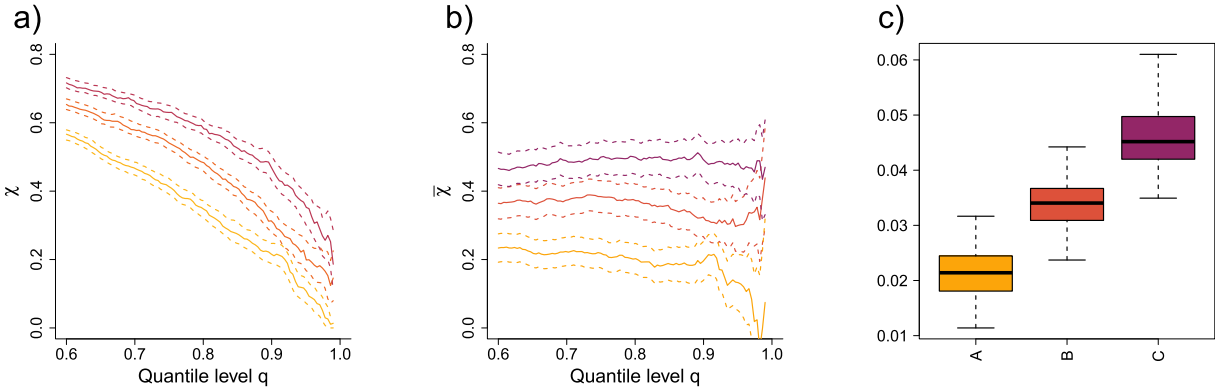

Fig. 6. (a–b) For each mediod of the clustering with $K = 3$, the estimated χ (a) and $\overline{\chi}$ (b) coefficient as a function of the quantile level q. Dashed lines highlight the 95% confidence bands. Colors correspond to the medoids illustrated in Fig. 5. (c) Boxplots of the estimated probabilities that both variables exceed their respective 0.9-quantile for the locations belonging to each cluster illustrated in Fig. 5. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Hillier and Dixon, 2020). Storms primarily arrive from the west or south west, leading to a high dependence between precipitation and wind speed whereas further eastward this dependence is strongly reduced.

Finally, in Fig. 6c we consider, for all the locations belonging to each of the three clusters, the joint probability $\mathbb{P}\left(X_1 > q_{0.9}^{X_1}, X_2 > q_{0.9}^{X_2}\right)$ that both variables X_1 and X_2 exceed their respective 0.9-quantiles $q_{0.9}^{X_1}$ and $q_{0.9}^{X_2}$. If X_1 and X_2 were independent, this probability should be roughly equal to $0.1^2 = 0.01$. For all three clusters, the observed median probability is substantially bigger than the independent case. Furthermore, the three clusters are characterized by different exceedances probabilities thus dividing the considered region in three area with milder or stronger risk of joint extremes.

4. Conclusions and outlook

Clustering spatial points is of high interest because it permits a deeper understanding of the hidden spatial structure of a physical phenomenon of interest, to improve the precision of statistical estimates in a context of scarce data and to identify areas in which common preventive actions can be useful in mitigating the effect of weather hazards.

So far, most research on spatial clustering of climate variables has focused on metrics based on a single variable of interest and its univariate distributional properties. To account for multivariate dependencies in the tails of the distribution, novel metrics to cluster spatial locations based on the dependence structures of multiple variables are required. Here we propose a clustering procedure capable to group spatial locations based on the bivariate tail dependence structure between two variables.

We apply the approach to compound precipitation and wind extremes over Ireland and Great Britain. From a first exploratory analysis, we found that the dependence in the extremes between these two variables is stronger in the western region. This is consistent with previous studies on the relationship between heavy precipitations and wind speed anomalies in this domain, and is due to the fact that extratropical cyclones mostly arrive from the west or south west. The clustering obtained with the proposed algorithm successfully divides the area of interest in zones characterized by a similar dependence structure in the extremes.

We have shown how the introduced metric d_{ij} can be used to cluster points into regions with similar tail dependence behaviour between two variables. A similar approach could be applied to cluster other types of compound extremes, for instance concurrent drought and heat (Zscheischler and Seneviratne, 2017; Manning et al., 2019), or to partition coasts into areas with similar compound flooding risk, for instance based on precipitation and storm surge extremes (Zheng et al., 2013; Bevacqua et al., 2019) or runoff and storm surge extremes (Ward et al., 2018; Couasnon et al., 2020).

The methodology could be extended to the multivariate case with $p > 2$ by building an overall divergence d_{ij} that considers all the pairwise cases, for example through averaging, or by directly defining the disjoint sets $A_1^{(j)}, \ldots, A_W^{(j)}$ in the $p-$ dimensional space using rules similar to the ones described in section 2.

The proposed clustering approach could also be used to evaluate how well atmospheric models are able to simulate coherent dependence patterns across space between extremes in different variables, similarly to the work of Bador et al. (2015) for univariate extremes. Overall, evaluating climate models with respect to compound events is important for gaining confidence in future climate model projections of such events (Zscheischler et al., 2021).

Testing whether and how dependence patterns of specific hazard combinations will change under warmer temperatures is another relevant application. Changes in dependence structure can substantially modify the risks associated with compound events and have already been identified for some hazard combinations such as heavy precipitation and extreme storm surge at US costs (Wahl et al., 2015), as well as hot and dry summers (Zscheischler and Seneviratne, 2017) in response to climate change. In this context, the proposed clustering method could be used by looking at the changes in the partitions obtained from different time slices representing warmer and colder periods. Another approach could be to weight the data points in equation (6) by a covariate associated with climate change, such as average global temperature.

Finally, it is important to understand which underlying factors determine the similarity of two distinct locations concerning their tail dependence structure. This question could be addressed with novel approaches from dynamical systems theory that allow to identify the atmospheric drivers behind compound extremes (De Luca et al., 2020). Thus, our bivariate divergence measure tailored to extremes is applicable to a variety of research questions in the emerging field of compound events.

Data availability statement

We acknowledge the ERA5 dataset (Copernicus Climate Change Service (C3S), 2017) produced by the European Center for Medium-Range Weather Forecasts (ECMWF, www.ecmwf.int/). The data are available from the website of the Copernicus Climate Change Service (cds.climate.copernicus.eu/).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the European COST Action DAMOCLES (CA17109). EV, SE and JZ acknowledge funding from the Swiss National Science Foundation (Doc.Mobility Grant 188229, Eccellenza Grant 186858 and Ambizione Grant 179876, respectively). JZ further acknowledges the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, Grant Agreement VH-NG-1537).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wace.2021.100318.

References

Asadi, P., Engelke, S., Davison, A.C., 2018. Optimal regionalization of extreme value distributions for flood estimation. J. Hydrol. 556, 182–193.
Bador, M., Naveau, P., Gilleland, E., Castellà, M., Arivelo, T., 2015. Spatial clustering of summer temperature maxima from the CNRM-CM5 climate model ensembles & E-OBS over Europe. Weather and Climate Extremes 9, 17–24.
Baker, L., Shaffrey, L., Scaife, A.A., 2018. Improved seasonal prediction of UK regional precipitation using atmospheric circulation. Int. J. Climatol. 38, e437–e453.
Bengtsson, L., Hodges, K.I., Keenlyside, N., 2009. Will extratropical storms intensify in a warmer climate? J. Clim. 22 (9), 2276–2301.
Bernard, E., Naveau, P., Vrac, M., Mestre, O., 2013. Clustering of maxima: spatial dependencies among heavy rainfall in France. J. Clim. 26 (20), 7929–7937.
Bevacqua, E., Maraun, D., Vousdoukas, M.I., Voukouvalas, E., Vrac, M., Mentaschi, L., Widmann, M., 2019. Higher probability of compound flooding from precipitation and storm surge in europe under anthropogenic climate change. Science Advances 5 (9). https://doi.org/10.1126/sciadv.aaw5531.
Copernicus Climate Change Service (C3S), 2017. ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global Climate, Copernicus Climate Change Service Climate Data Store (CDS).
Couasnon, A., Eilander, D., Muis, S., Veldkamp, T., Haigh, I., Wahl, T., Winsemius, H., Ward, P., 2020. Measuring compound flood potential from river discharge and storm surge extremes at the global scale. Nat. Hazards Earth Syst. Sci. 20 (2), 489–504.
De Luca, P., Messori, G., Pons, F.M.E., Faranda, D., 2020. Dynamical systems theory sheds new light on compound climate extremes in Europe and Eastern North America. Q. J. R. Meteorol. Soc. 1–15. https://doi.org/10.1002/qj.3757.
Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., et al., 2011. The era-interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137 (656), 553–597.
Durante, F., Pappadà, R., Torelli, N., 2015. Clustering of time series via non-parametric tail dependence estimation. Stat. Pap. 56 (3), 701–721.
Engelke, S., Naveau, P., Zhou, C., 2019. Kullback-Leibler divergence for multivariate extremes. In: Presented at ISI World Statistics Congress in Kuala Lumpur. In Preparation..
Field, C.B., Barros, V., Stocker, T.F., Dahe, Q., 2012. Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Fowler, H., Ekström, M., Kilsby, C., Jones, P., 2005. New estimates of future changes in extreme rainfall across the UK using regional climate model integrations. 1. Assessment of control climate. J. Hydrol. 300 (1–4), 212–233.
Gouldby, B., Wyncoll, D., Panzeri, M., Franklin, M., Hunt, T., Hames, D., Tozer, N., Hawkes, P., Dornbusch, U., Pullen, T., 2017. Multivariate extreme value modelling of sea conditions around the coast of england. In: Proceedings of the Institution of Civil Engineers-Maritime Engineering, vol. 170. Thomas Telford Ltd, pp. 3–20.
Grimaldi, S., Petroselli, A., Salvadori, G., De Michele, C., 2016. Catchment compatibility via copulas: a non-parametric study of the dependence structures of hydrological responses. Adv. Water Resour. (90), 116–133.
Heffernan, J.E., Tawn, J.A., 2004. A conditional approach for multivariate extreme values (with discussion). J. Roy. Stat. Soc. B 66 (3), 497–546.
Hendry, A., Haigh, I.D., Nicholls, R.J., Winter, H., Neal, R., Wahl, T., Joly-Laugel, A., Darby, S.E., 2019. Assessing the characteristics and drivers of compound flooding events around the UK coast. Hydrol. Earth Syst. Sci. 23 (7), 3117–3139. https://doi.org/10.5194/hess-23-3117-2019.

Hillier, J.K., Dixon, R.S., 2020. Seasonal impact-based mapping of compound hazards. Environ. Res. Lett. 15 (11), 114013. https://doi.org/10.1088/1748-9326/abbc3d.
Jonathan, P., Ewans, K., Randell, D., 2014. Non-stationary conditional extremes of northern north sea storm characteristics. Environmetrics 25 (3), 172–188.
Kaufman, L., Rousseeuw, P.J., 2009. Finding Groups in Data: an Introduction to Cluster Analysis, vol. 344. John Wiley & Sons.
Kolev, N., Anjos, U.d., Copulas, B. V. d. M. Mendes, 2006. A review and recent developments. Stoch. Model 22 (4), 617–660.
Ledford, A.W., Tawn, J.A., 1996. Statistics for near independence in multivariate extreme values. Biometrika 83 (1), 169–187.
Ledford, A.W., Tawn, J.A., 1997. Modelling dependence within joint tail regions. J. Roy. Stat. Soc. B 59 (2), 475–499.
Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., Risbey, J., Schuster, S., Jakob, D., Stafford-Smith, M., 2014. A compound event framework for understanding extreme impacts. Wiley Interdisciplinary Reviews: Climate Change 5 (1), 113–128.
Liberato, M.L., 2014. The 19 January 2013 windstorm over the North Atlantic: large-scale dynamics and impacts on Iberia. Weather and Climate Extremes 5, 16–28.
Mailier, P.J., Stephenson, D.B., Ferro, C.A.T., Hodges, K.I., 2006. Serial clustering of extratropical cyclones. Mon. Weather Rev. 134 (8), 2224–2240. https://doi.org/10.1175/MWR3160.1.
Manning, C., Widmann, M., Bevacqua, E., Loon, A.F.V., Maraun, D., Vrac, M., aug 2019. Increased probability of compound long-duration dry and hot events in europe during summer (1950–2013). Environ. Res. Lett. 14 (9), 094006 https://doi.org/10.1088/1748-9326/ab23bf.
Martius, O., Pfahl, S., Chevalier, C., 2016. A global quantification of compound precipitation and wind extremes. Geophys. Res. Lett. 43 (14), 7709–7717.
Pappadà, R., Durante, F., Salvadori, G., De Michele, C., 2018. Clustering of concurrent flood risks via hazard scenarios. Spatial Statistics 23, 124–142.
Poon, S.-H., Rockinger, M., Tawn, J., 2004. Extreme value dependence in financial markets: diagnostics, models, and financial implications. Rev. Financ. Stud. 17 (2), 581–610.
Raveh-Rubin, S., Wernli, H., 2015. Large-scale wind and precipitation extremes in the Mediterranean: a climatological analysis for 1979–2012. Q. J. R. Meteorol. Soc. 141 (691), 2404–2417.
Ribeiro, A.F.S., Russo, A., Gouveia, C.M., Páscoa, P., Zscheischler, J., 2020. Risk of crop failure due to compound dry and hot extremes estimated with nested copulas. Biogeosciences 17 (19), 4815–4830. https://doi.org/10.5194/bg-17-4815-2020.
Rohrbeck, C., Tawn, J.A., 2020. Bayesian spatial clustering of extremal behaviour for hydrological variables. J. Comput. Graph Stat. 1–15.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Saunders, K.R., Stephenson, A.G., Karoly, D.J., 2020. A regionalisation approach for rainfall based on extremal dependence. Extremes. https://doi.org/10.1007/s10687-020-00395-y.
Sibuya, M., 1960. Bivariate extreme statistics, I. Ann. Inst. Stat. Math. 11 (2), 195–210.
Svensson, C., Brookshaw, A., Scaife, A., Bell, V., Mackay, J., Jackson, C., Hannaford, J., Davies, H., Arribas, A., Stanley, S., 2015. Long-range forecasts of UK winter hydrology. Environ. Res. Lett. 10 (6), 064006.
Wadsworth, J.L., Tawn, J.A., 2012. Dependence modelling for spatial extremes. Biometrika 99 (2), 253–272.
Wahl, T., Jain, S., Bender, J., Meyers, S.D., Luther, M.E., 2015. Increasing risk of compound flooding from storm surge and rainfall for major US cities. Nat. Clim. Change 5 (12), 1093–1097.
Ward, P.J., Couasnon, A., Eilander, D., Haigh, I.D., Hendry, A., Muis, S., Veldkamp, T.I. E., Winsemius, H.C., Wahl, T., 2018. Dependence between high sea-level and high river discharge increases flood hazard in global deltas and estuaries. Environ. Res. Lett. 13 (8), 084012 https://doi.org/10.1088/1748-9326/aad400.
Weston, K.J., Roy, M.G., 1994. The directional-dependence of the enhancement of rainfall over complex orography. Meteorol. Appl. 1 (3), 267–275.
Winter, H.C., Tawn, J.A., Brown, S.J., et al., 2016. Modelling the effect of the el nino-southern oscillation on extreme spatial temperature events over Australia. Ann. Appl. Stat. 10 (4), 2075–2101.
Zheng, F., Westra, S., Sisson, S.A., 2013. Quantifying the dependence between extreme rainfall and storm surge in the coastal zone. J. Hydrol. 505, 172–187.
Zscheischler, J., Naveau, P., Martius, O., Engelke, S., Raible, C.C., 2021. Evaluating the dependence structure of compound precipitation and wind speed extremes. Earth Syst. Dynam. 12 (1), 1–16.
Zscheischler, J., Seneviratne, S.I., 2017. Dependence of drivers affects risks associated with compound events. Science Advances 3 (6), e1700263.
Zscheischler, J., Mahecha, M.D., Harmeling, S., 2012. Climate classifications: the value of unsupervised clustering. Procedia Computer Science 9, 897–906. https://doi.org/10.1016/j.procs.2012.04.096.
Zscheischler, J., Michalak, A.M., Schwalm, C., Mahecha, M.D., Huntzinger, D.N., Reichstein, M., Berthier, G., Ciais, P., Cook, R.B., El-Masri, B., Huang, M., Ito, A., Jain, A., King, A., Lei, H., Lu, C., Mao, J., Peng, S., Poulter, B., Ricciuto, D., Shi, X., Tao, B., Tian, H., Viovy, N., Wang, W., Wei, Y., Yang, J., Zeng, N., 2014. Impact of large-scale climate extremes on biospheric carbon fluxes: an intercomparison based on MsTMIP data. Global Biogeochem. Cycles 28, 585–600. https://doi.org/10.1002/2014GB004826.
Zscheischler, J., Westra, S., Van Den Hurk, B., Seneviratne, S., Ward, P., Pitman, A., AghaKouchak, A., Bresch, D., Leonard, M., Wahl, T., et al., 2018. Future climate risk from compound events. Nat. Clim. Change 8 (6), 469–477.
Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Horton R.M., C.R., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M.D., Maraun, D., Ramos, A.M., Ridder, N., Thiery, W., Vignotto, E., 2020. A typology of compound weather and climate events. Nature Reviews Earth & Environment 1 (7), 333–347. https://doi.org/10.1038/s43017-020-0060-z.